Cleaning the Europarl Corpus for Linguistic Applications

Johannes Graën **Dolores Batinic** **Martin Volk**
Institute of Computational Linguistics
University of Zurich, Zurich, Switzerland
{graen|batinic|volk}@cl.uzh.ch

Abstract

We discovered several recurring errors in the current version of the Europarl Corpus originating both from the web site of the European Parliament and the corpus compilation based thereon. The most frequent error was incompletely extracted metadata leaving non-textual fragments within the textual parts of the corpus files. This is, on average, the case for every second speaker change.

We not only cleaned the Europarl Corpus by correcting several kinds of errors, but also aligned the speakers' contributions of all available languages and compiled everything into a new XML-structured corpus. This facilitates a more sophisticated selection of data, e.g. querying the corpus for speeches by speakers of a particular political group or in particular language combinations.

1 Introduction

Koehn (2005) first presented his compilation of the Europarl Corpus comprising the European Parliament's debates in 2001 and has continued updating it with the latest available data since then. For compiling the transcriptions of the debates into a corpus, he downloaded the particular web pages for any available language[1] from the Parliament's web site[2]. He then parsed the HTML source code in order to separate markup, meta-information, like the structure of the debates (chapters and turns), speaker information and so forth, as well as actual textual data (i.e. topics, comments and the respective speakers' speeches) and transferred the latter two into plain text files. As of today, the corpus has grown to 968 plenary sessions from 1996 to 2011 in up to 21 languages in parallel[3].

The plenary sessions consist of a list of agenda items (called 'chapters') themselves consisting of speech contributions of one or more speakers in combination with descriptive comments by the transcribers (hereinafter called 'turns'). Usually, the first and last turn within a chapter lies with the president of the European Parliament.

The Europarl Corpus is widely used for many diverse language technology applications such as "word sense disambiguation, anaphora resolution, information extraction" (ibid.), statistical machine translation (Cohn and Lapata 2007), grammar projection (Bouma et al. 2008) "unsupervised part-of-speech tagging" (Das and Petrov 2011) or "learning multilingual semantic representations" (Hermann and Blunsom 2014).

When we evaluated the appropriateness of the Europarl Corpus for our intended applications (namely part-of-speech tagging, chunking and parsing as well as word, chunk and tree alignments over parallel texts), we encountered sev-

[1] Before 2004, the European Union had 11 official languages which are part of the first Europarl Corpus version.

[2] Available at http://www.europarl.europa.eu/.

[3] Although Maltese and Irish have become official languages as of 2004 and 2007 respectively, no transcripts of the debates are available in these languages.

eral problems which might be negligible when the data is analyzed statistically, but impair the results with regard to individual examples. We found that shallow cleaning already led to a considerable improvement on further processing steps like part-of-speech tagging or turn alignment. The latter is necessary before sentence alignment due to partial absence of translations that would evoke wrong sentence alignments in most cases. That is why we decided to clean Koehn's corpus for our own purposes and with the objective of making the result publicly available again.

Crawling the web pages of the European Parliament's debates and extracting the aforesaid information by ourselves could have been an alternative way to obtain clean data. As some of the errors originate from the European Parliament's website and thus need to corrected anyway and the debates before the parliamentary term of 1999–2004 are no longer available online, we opted for cleaning the existing corpus data.

For publishing our corpus compilation, we decided to store the cleaned version of the Europarl Corpus in XML format, which will enable a more fine-grained selection of data than the original plain text files. By means of XPath (Clark, DeRose, et al. 1999), one can query the corpus for particular speakers, dates or political groups. Building a sub-corpus of transcribed speeches in one language with translations to a second one and comparing it to another one where transcriptions and translations are arranged the other way round is just as simple as using XPath expressions to match the language code of a speaker's contribution and the language label of the text. Additionally, it is possible to address the structure of the respective discourse (from sequential comments of each group's position to dialogue alike controversies).

2 Error classification

Koehn refers to the extraction of textual data from the website of the European Parliament as a "cumbersome enterprise". Hence, several formatting problems such as the diversified use of encoding alternatives for certain characters, HTML entities or wrong quotation marks (see Ex. 1) made it into his published corpus.

(1)
1. Non sono contrario a Eurojust, ma non deve trasformarsi in una "super-istituzione".
2. Je l'appellerais un vote d'»avis conforme élargi», qui n'est pas …
3. … und zwar wegen der Existenz so genannter ,,Energieinseln" wie dem Ostseeraum, …

In addition to that, numerous errors were introduced by Koehn's corpus compilation. On the one hand, speaker information was often incompletely extracted from the web pages, i.e. meta-information such as the language used by the speaker is classified as part of the speakers' utterance (see Ex. 2), or comprises textual information, mostly comments structuring the transcriptions (see Ex. 3).

(2)
1. (RO) Бих искала да поздравя г-н Stolojan за …
2. , Kommissionen. (EN) Når det gælder …
3. Miller (PSE). (EN) Herr Präsident, ich …
4.). (IT) Κυρία Πρόεδρε, …
5. (RO) Бих искала да поздравя г-н Stolojan за …

(3)
1. <SPEAKER ID="66" LANGUAGE="PL" NAME="Protasiewicz (PPE-DE)." AFFILIATION="(Applaus)">
2. <SPEAKER ID="11" LANGUAGE="" NAME="Tannock (PPE-DE)." AFFILIATION="(">
3. <SPEAKER ID="115" LANGUAGE="" NAME="" AFFILIATION="The Minutes of the previous sitting were approved.)"/>

On the other hand, Koehn applied shallow tokenization rules, which in some languages resulted in partly tokenized texts. When apostrophized prepositional articles in Italian and French are separated from the following word with a white space ("all' uomo" instead of "all'uomo") and the text is fed to a tokenizer afterwards, for instance, the already tokenized parts will be handled again, thus leading to potential erroneous output as shown in Figure 1.

A more severe problem is that text in HTML tags within the textual parts of the pages is omitted (Example 4 shows a sample sentence from the website of the European Parliament (1) and its counterpart in Koehn's corpus compilation (2)) and the original text is thus unrecoverable.

(4)
1. Il caso Terni è, per molti versi, la punta di un iceberg.
2. Il caso Terni è, per molti versi, la punta di un .

Apart from the errors introduced by Koehn, we identified several problems originating from the text of the original web pages. Besides certain

1. L' ordine del giorno reca la fissazione dell' ordine dei lavori.
   /NUM '/PON ordine/NOM del/PRE:det giorno/NOM recere/VER:cpre il/DET:def fissazione/NOM <unknown>/NOM '/PON ordine/NOM del/PRE:det lavorio|lavoro/NOM ./SENT
2. L'ordine del giorno reca la fissazione dell'ordine dei lavori.
   il/DET:def ordine/NOM del/PRE:det giorno/NOM recere/VER:cpre il/DET:def fissazione/NOM del/PRE:det ordine/NOM del/PRE:det lavorio|lavoro/NOM ./SENT

Figure 1: Example sentence with corresponding output of the TreeTagger (cf. Schmid 1994) which performs tokenization before tagging. (1) shows the the sentence as it appears in Koehn's corpus, in (2) the partial tokenization is undone. Relevant corresponding parts are highlighted, the correct tagging is underlined.

parts being absent in otherwise completely translated documents, comments were often not translated (see Ex. 5). While this kind of missing data cannot be adjusted at all, correction candidates include, for example, wrong punctuation (especially quotation marks), the common misspelling of è in Italian as e' and perchè instead of perché (The same applies to ché and affinché), missing space characters in front of French punctuation signs and wrong number formats.

(5)
1. Schluss der Sitzung
   <P>
   (The sitting closed at 22.25)

We classified the errors and problem as shown in Figure 2 according to their source, impact as well as frequency and grouped them into categories which best describe their nature. The impact of a particular type of error is classified as "low", "medium" or "high", depending on how further processing tools are affected by the particular error type, i.e. whether they skip or autocorrect it or produce wrong output or analysis. We evaluated the output of tools for common processes such as tokenization, part-of-speech tagging or parsing in order to decide what impact a particular error type might have. However, it will vary for different kind of applications.

3 Correcting errors and enriching the corpus

In our cleaning of the corpus, we traversed all of Koehn's corpus files for each plenary session and extracted structural elements, meta-information, comments and either original (transcribed) or translated speeches. According to the type of data, we used different cleaning rules.

Language specifications not belonging to the official or semi-official languages of the European Union, for example, are in most cases obvious mistakes (e.g. using uk (Ukrainian) or gb (not assigned) as language code instead of en (English) for a British speaker) and not taken over to the cleaned corpus so that the respective turns lack the attribute for the original language of the utterance. In order to enable the speaker turn alignment, we also identified the utterances of the chairmen of the parliament (see Ex. 6 for examples) for each language by comparing their names to the respective language's term variants for the president (in German, for instance: "Der Präsident", "Die Präsidentin", "Präsident" and "Präsidentin").

(6)
1. <SPEAKER ID="213" NAME="Le Président">
2. <SPEAKER ID="213" NAME="Πρόεδρος">
3. <SPEAKER ID="213" NAME="elnök">
4. <SPEAKER ID="213" NAME="De Voorzitter">
5. <SPEAKER ID="213" NAME="Talmannen">

In addition, we split multiple speaker names like "Bonde, Lis Jensen i Sandbæk" (taken from the Romanian text) into the particular names ("Bonde", "Lis Jensen" and "Sandbæk") and marked the turn as having multiple authors. Multiple authors are only possible in written parts added to the transcripts, usually being the "explanations of vote", in order to facilitate the interlingual alignment of turns and the access to the author information in the corpus.

Having removed the meta-information, we applied a set of further cleaning rules to the actual textual information. This set comprises corrections for wrong characters and punctuation, marks URLs, parliamentary reports as well as legislative procedures and undoes the partly performed tokenization on a per language level. We also applied a multitude of rules to identify all kinds of correct and wrong quotation marks and unified them.[4] Additional language-specific rules help us to meet orthography requirements.

[4] We provide the correct language-specific use of quotation marks in the respective language as additional informa-

category	error/problem	source[5]	impact	frequency[6]
coding	invalid UTF-8 encoding	K	low	$< 10^{-5}$/files
	undecoded HTML entities	EP	medium	$< 10^{-2}$/lines
	code variants[7]	EP	medium	$> 6\,\%$/lines
orthography	consistently misspelled words	EP	medium	$< 2\,\%$/lines[8]
	wrong/incoherent quotation marks (see Ex. 1)	both	low	$< 1\,\%$/lines
missing data	words omitted (see Ex. 4)	K	high	$> 10^{-3}$/lines
	comments untranslated (see Ex. 5)	EP	low	$> 10^{-5}$/lines
	non-matching turns[9]	both	high	$> 10^{-3}$/turns
processing	text partly tokenized	K	medium	$> 10^{-3}$/tokens
	text marked as meta-information (see Ex. 3)	K	low	$> 1\,\%$/lines
	meta-information marked as text (see Ex. 2)	K	high	$> 6\,\%$/lines

Figure 2: Error classification scheme.

After the corresponding documents for a particular plenary session in any available language have been mapped to internal representations, we aligned the corresponding speakers' turns in all languages. In the majority of cases (58%), the respective documents have exactly the same structure so that the alignment process is straightforward. For all other cases, we searched for possible alignments with respect to the given order of turns and calculated a score based on the Levenshtein distance between two speaker names, the property of a speaker being president of the European Parliament or not, the chapter a turn is listed in and the count of textual parts within that turn. We then computed a list of alignments minimizing that measure. In this vein, we are able to correct wrongly aligned turns, i.e. for instance those that were only based on the id attribute given by Koehn (2005).

tion so that the text with markup can instantaneously be converted to its correct plain form.

[5]Either the Website of the European Parliament (EP) or Koehn's compilation (K) or both.

[6]The frequencies are calculated or estimated based on the source corpus files, their lines (text or meta-information) or tokens.

[7]Various hyphens and dashes as well as homoglyphs.

[8]Calculated for misspelled è in Italian. As this is the most frequent case of misspelling and we found that approximately 2 % of the lines of the Italian texts, the overall frequency needs to be lower.

[9]The number of turns of the respective languages in a particular chapter or session don't match. This can be due to a wrong classification of text as meta-information, meta-information as text or the absence of one or more turns.

4 Evaluation

We calculated that at least 6% of all lines from Koehn's corpus erroneously contain meta-information, which we were able to correct. Sole language information was the most frequent case (cf. the last example of Fig. 2). This quantity corresponds to 50% of the meta-data definitions in the corpus, thus implying that in half of the cases Koehn's meta-data extraction rules failed.

About 1% of the text lines in Koehn's Europarl contain comments introduced by the transcriber of the sessions. We marked all these comments, even some that did not possess any specific formatting, by comparing lines with a handcrafted list. In about 2% of the lines we were able to detect and eliminate non textual fragments originating from the European Parliament's web pages.

We found quotes in 3% of the lines which we marked as such by applying rules that were made to recognize even the wrong quotation marks (cf. Fig. 1).

Since we created restrictive rules to only correct these systematic errors that we identified, none of them remain uncorrected and a quantitative evaluation of corrected errors would be futile. Nonetheless, 12% of all apostrophized articles, prepositions and prepositional articles, for instance, features a following white space (probably due to Koehn's shallow tokenization rules) which is prone to cause problems as depicted in Figure 1.

Some errors known to us, including, for instance, Catalan text within the Spanish parts or

turns in one language located in a different chapter than the corresponding parts of the other languages and hence not being aligned, occur only infrequently. Thus, we decided to leave them alone.

5 Conclusions

We improved the quality of the Europarl Corpus described by Koehn (ibid.) and recompiled in 2012 by

- correcting the classification of meta-data versus textual data,
- unifying punctuation marks and other kinds of character classes,
- marking identifiers of political groups in the European Parliament and
- removing fragments of characters which are legacy of the original source but do not belong to the textual data.

Furthermore, we enriched the corpus by

- marking agenda items, comments and speech parts where distinguishable,
- marking quotes in a way that they can be converted to each language's preferences, and
- aligning speakers' contributions (turns) in all available languages.

We considered adding the respective speakers' mother tongue as supplementary information in order to pave the way for an even deeper linguistic investigation. Unfortunately, we did not find any data that could have enabled us to do so with a reasonable effort and accuracy.

The resulting edited and recompiled corpus serves (like Koehn's original one) as a rich source for any kind of linguistic application, but in addition to that provides easier access to cleaned text and meta-data both being arranged in an XML structure (see Listing 1). Corresponding speaker turns (one being the speech's transcription and the others translation of it) are aligned by sharing the same turn node.

The corrected and structured Europarl Corpus as well as some technical documentation can be obtained from `http://pub.cl.uzh.ch/purl/costep`.

6 Future work

The debates of the European Parliament keep being an important resource of parallel texts in many languages for a multitude of language technology applications. New web pages are added for every completed plenary sessions and made available in an increasing number of languages.

We place great importance on the availability of up-to-date, turn-aligned parallel texts from the European Parliament's debates and suggest to integrate the tasks of crawling the web pages, cleaning the textual data and aligning the respective speaker's turns. We believe that in this vein certain types of errors can be corrected with less effort while others can be entirely avoided.

Acknowledgment

This research was supported by the Swiss National Science Foundation under grant 105215_146781/1 through the project "SPARCLING – Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation".

References

Bouma, Gerlof et al. (2008). "Parallel LFG Grammars on Parallel Corpora: A base for practical triangulation". In: *Proceedings of the Lexical Functional Grammar (LFG) Conference*. (Sydney). International Lexical Functional Grammar Association (ILFGA), pp. 169–189.

Clark, James, Steve DeRose, et al. (1999). *XML path language (XPath) version 1.0*.

Cohn, Trevor and Mirella Lapata (2007). "Machine translation by triangulation: Making effective use of multi-parallel corpora". In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. (Prague). Vol. 45. 1. Association for Computational Linguistics (ACL), pp. 728–735.

Das, Dipanjan and Slav Petrov (2011). "Unsupervised Part-of-speech Tagging with Bilingual Graph-based Projections". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies - Volume 1*. (Portland), pp. 600–609.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<session date="2010-05-05">
  <chapter id="1">
    <headline language="bg">Възобновяване на сесията</headline>
    <headline language="cs">Pokračování zasedání</headline>
    <headline language="da">Genoptagelse af sessionen</headline>
    <headline language="de">Wiederaufnahme der Sitzungsperiode</headline>
    <headline language="el">Επανάληψη της συνόδου</headline>
    <headline language="en">Resumption of the session</headline>
    <headline language="es">Reanudación del período de sesiones</headline>
    <headline language="et">Istungjärgu jätkamine</headline>
    <headline language="fi">Istuntokauden uudelleen avaaminen</headline>
    <headline language="fr">Reprise de la session</headline>
    <headline language="hu">Az ülésszak folytatása</headline>
    <headline language="it">Ripresa della sessione</headline>
    <headline language="lt">Sesijos atnaujinimas</headline>
    <headline language="lv">Sesijas atsākšana</headline>
    <headline language="nl">Hervatting van de zitting</headline>
    <headline language="pl">Wznowienie sesji</headline>
    <headline language="pt">Reinício da sessão</headline>
    <headline language="ro">Reluarea sesiunii</headline>
    <headline language="sk">Pokračovanie prerušeného zasadania</headline>
    <headline language="sl">Nadaljevanje zasedanja</headline>
    <headline language="sv">Återupptagande av sessionen</headline>
    <turn id="1">
      <speaker president="yes" language="el">
        <text language="bg">
          <p type="speech">Възобновявам сесията на Европейския парламент, прекъсната на 22 април 2010 г.</p>
          <p type="speech">Протоколът от 22 април 2010 г. беше раздаден.</p>
          <p type="speech">Има ли някакви коментари?</p>
          <p type="comment">Протоколът от предишното заседание е одобрен</p>
        </text>
        <text language="cs">
          <p type="speech">Prohlašuji přerušené zasedání Evropského parlamentu ze dne 22. dubna 2010 za obnovené.</p>
          <p type="speech">Zápis z jednání ze dne 22. dubna 2010 byl rozdán.</p>
          <p type="speech">Má někdo připomínky?</p>
          <p type="comment">Zápis z předchozího zasedání byl schválen</p>
        </text>
        <text language="da">
          <p type="speech">Jeg erklærer Europa-Parlamentets session, der blev afbrudt torsdag den 22. april 2010, for genoptaget.</p>
          <p type="speech">Protokollen fra mødet den 22. april 2010 er omdelt.</p>
          <p type="speech">Hvis ingen gør indsigelse, betragter jeg den som godkendt.</p>
          <p type="comment">Protokollen fra foregående møde godkendtes</p>
        </text>
        <text language="de">
          <p type="speech">Ich erkläre die am 22. April 2010 unterbrochene Sitzung des Europäischen Parlaments für wieder aufgenommen.</p>
          <p type="speech">Das Protokoll vom 22. April 2010 wurde ausgeteilt.</p>
          <p type="speech">Gibt es dazu Anmerkungen?</p>
          <p type="comment">Das Protokoll der vorherigen Sitzung wird angenommen</p>
        </text>
        <text language="el">
          <p type="speech">Κηρύσσω την επανάληψη της συνόδου του Ευρωπαϊκού Κοινοβουλίου η οποία είχε διακοπεί στις 22 Απριλίου 2010.</p>
          <p type="speech">Τα Συνοπτικά Πρακτικά της συνεδρίασης της 22ας Απριλίου 2010 έχουν διανεμηθεί.</p>
          <p type="speech">Υπάρχουν παρατηρήσεις επ' αυτών;</p>
          <p type="comment">Εγκρίνονται τα Συνοπτικά Πρακτικά της προηγούμενης συνεδρίασης</p>
        </text>
        <text language="en">
          <p type="speech">I declare resumed the session of the European Parliament adjourned on 22 April 2010.</p>
          <p type="speech">The Minutes of 22 April 2010 have been distributed.</p>
          <p type="speech">Are there any comments?</p>
          <p type="comment">The Minutes of the previous sitting were approved</p>
        </text>
```

Listing 1: Excerpt from a turn-aligned XML corpus file for a particular plenary session.

Hermann, Karl Moritz and Phil Blunsom (2014). "Multilingual Models for Compositional Distributed Semantics". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. (Baltimore). Association for Computational Linguistics (ACL).

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Machine Translation Summit*. (Phuket). Vol. 5. Asia-Pacific Association for Machine Translation (AAMT), pp. 79–86.

Schmid, Helmut (1994). "Probabilistic part-of-speech tagging using decision trees". In: *Proceedings of International Conference on New Methods in Natural Language Processing (NeMLaP)*. (Manchester). Vol. 12, pp. 44–49.